Information Processing and Management 60 (2023) 103254

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A multimodal time-series method for gifting prediction in live streaming platforms

Dinghao Xi [a], Liumin Tang [a], Runyu Chen [b], Wei Xu [a,*]

[a] School of Information, Renmin University of China, Beijing 100872, PR China
[b] School of Information Technology and Management, University of International Business and Economics, Beijing 100029, PR China

ARTICLE INFO

Keywords:
Live streaming
Gifting prediction
Multimodal fusion
Transformer

ABSTRACT

Viewer gifting is an important business mode in live streaming industry, which closely relates to the income of the platforms and streamers. Previous studies on gifting prediction are often limited to cross-section data and consider the problem from the macro perspective of the whole live streaming. However, the multimodal information and the time accumulation effect of live streaming content on viewer gifting behavior are ignored. In this paper, we put forward a multimodal time-series method (MTM) for predicting real-time gifting. The core module of the method is the multimodal time-series analysis (MTA), which targets at effectively fusing multi-modal information. Specifically, the proposed orthogonal projection (OP) model can promote cross-modal information interaction without introducing additional parameters. To achieve the interaction of multi-modal information at the same level, we also design a stackable joint representation layer, which makes each target modality's representation (visual, acoustic and textual modality) can benefit from all the other modalities. The residual connections are introduced as well to ensure the integration of low-level and high-level information. On our dataset, our model shows improved performance compared to other advanced models by at least 8% on F1. Meanwhile, the MTA is able to meet the real-time requirements of the live streaming setting, and has demonstrated its robustness and transferability in other tasks. Our research may offer some insights about how to efficiently fuse multimodal information, and contribute to the research on viewer gifting behavior prediction in the live streaming context.

1. Introduction

The past few years have witnessed the rise of live streaming over the world. Numerous social network giants, such as YouTube, Twitter and Facebook, have been providing live streaming services since 2011 (Lin et al., 2021). Twitch, which made it fortune by video game live streaming in 2011, has captured young people's hearts and now serves over 31 million daily active users[1]. In China, the number of live streaming users has reached 703 million by 2021, accounting for 68.2% of the overall citizens[2].

In the live streaming context, streamers create real-time content via the service provided by the platform. The ways they use to attract audiences include but are not limited to singing, chatting, playing games, sharing life experience and knowledge (Lu et al.,

* Corresponding author.
E-mail address: weixu@ruc.edu.cn (W. Xu).

[1] Twitch (2022). Audience. http://twitchadvertising.tv/audience/.
[2] CNNIC (2022). The 49th Statistical Report on Internet Development in China. https://www.cnnic.net.cn/n4/2022/0401/c88-1131.html.

https://doi.org/10.1016/j.ipm.2022.103254
Received 13 June 2022; Received in revised form 15 December 2022; Accepted 19 December 2022

2018). For viewers, as the consumers of live streaming content, they can participate in the live streaming by posting live comments, sending emojis, or hitting the like button. In order to motivate the streamers to create more content and interact more with the viewers, many platforms have carried out virtual gifting services. The viewers can pay for the virtual gifts and send them to support their favorite streamers. The live streaming platforms earn revenue from the virtual gifts selling, and streamers can get a percentage of the revenue (Wang et al., 2019). Viewer gifting, as a crucial business mode in live streaming economy (Zhou et al., 2019), makes great economic value for both the platforms and the streamers (Tu et al., 2018).

There exist numerous studies on viewer gifting, most of which focus on the motivation and the contributing factors of gifting behavior (Wohn et al., 2018; Li et al., 2021). For instance, Liu et al. (2022) proposed that viewer value perception can serve as a predictor of the number and the amount of the gifts given by viewers. However, when analyzing and predicting viewer gifting behavior, previous research has not made full use of the information available to the viewers, such as the live streaming content and live comments. Besides, these studies are limited to cross-section data, which macroscopically consider viewers' gifting from the perspective of the whole live and ignore the real-time nature of viewer gifting.

To address the above shortcomings, the research target of our study is to propose a method to predict the real-time gifting during the live streaming. Compared to previous studies, we comprehensively consider the influences of multimodal information (text, audio, and image) on gifting. Our prediction model is suitable for real-time prediction scenarios, which can help the live streaming platforms to timely identify potential streamers (who may receive more virtual gifts from viewers in the future) and thus to recommend potential streamers and increase economic returns.

In the online video and live streaming context, using multimodal information for prediction tasks has become the main stream routine (Xi et al., 2021). Previous research has proven the effect of live streaming scene, live comments, and the time of the live on viewers' gifting behavior (Lin et al., 2021; Xu et al., 2021). Thus, we include these multimodal signals for gifting prediction. The challenges of real-time prediction of the gifting lie in how to construct a multimodal model with high performance and low computational cost to meet the requirements of real-time prediction. In this research, we integrate the sequence information of textual, acoustic and visual information to build our multimodal time-series method (MTM) for gifting prediction. The core module of MTM is named multimodal time-series analysis (MTA). In MTA, we design a new multimodal fusion method called orthogonal projection (OP) model, which can be combined with the transformer encoder and effectively promote information interaction between modes. The OP model enables each mode to receive more useful information from other modes and reduces the redundancy of cross-modal information, generating better fused representations for the prediction task. Meanwhile, we also test four different architectures: early fusion, late fusion, hybrid fusion and hybrid residual fusion, and select the model with best performance as our final model. Based on the evaluation results on our dataset and on three standard datasets for different tasks, the hybrid residual fusion method distinguishes itself with stable and excellent performance. On our dataset with a 5 min observation window, our model outperformed other advanced models by at least 8% on F1. While on the dataset with a 10 min time window, our model leads by 16.54%.

Our contributions can be summarized as follows. First, we construct a multimodal time-series method (MTM) for gifting prediction during the live streaming. As the core part of the method, the proposed multimodal time-series analysis (MTA) module can handle textual, acoustic, and visual information effectively. Second, we design the orthogonal projection (OP) model in MTA, which reinforces a target modality using the fused information computed on the two other modalities. This technique reduces the method complexity while improving the performance. Third, we not only test the MTA on our dataset and get better performance, but also apply it to three standard datasets of different tasks. The results reveal that our model is robust enough to migrate to other task scenarios.

The rest of the paper is organized as follows. In Section 2, we introduce the related work. We describe our multimodal time-series method in detail in Section 3. In the following Section 4, we provide the data description and all the experimental results. In Section 5, we conclude our research and discuss the future work.

2. Related Work

2.1. Viewer gifting in live streaming

Previous research on viewer gifting in live streaming can be divided into two categories: one is to explore the motivations of viewers' gifting behavior, and the other focuses on the influencing factors of gifting. In live streaming, viewers have various motivations for gifting. For example, Wu et al. (2022) found that materialist beliefs and downstream self-related motivations could be the main driving factors of viewer gifting. Furthermore, some researchers look into the more specific motivations of gifting. Some viewers pay for their happy experiences or the knowledge they get from the streamer, while others hope that their gifting can help the streamers to sustain and improve their live streaming content. Finally, part of the viewers hopes to help solve offline social issues via gifting (Wohn et al., 2018).

A general conclusion can be drawn from all the relevant literature on viewer gifting is that the influencing factors of gifting are quite complex, including both external and internal factors. In terms of external factors, for example, Zhou et al. (2019) suggested that the viewer-viewer interaction (live comments) could affect viewers' arousal level, thereby motivating gifting behavior. Lin et al. (2021) found that the cheerful expression of streamers in live streaming would also inspire the viewers to pay for gifts. As for the internal reasons, both class identity and relational identity have been proven to be important contributing factors of viewer gifting (Li et al., 2021).

These studies mainly adopt traditional regression models, such as logistic regression (LR), panel vector autoregression (PVAR) and structural equation model (SEM), to explore the motivations and the influencing factors of viewer gifting. Moreover, these studies cannot include all the objective factors and relay heavily on the cross-sectional data, which ignore the time accumulation effect on

viewer gifting. Viewers' gifting behavior may be triggered by a previous period of time of the live streaming content rather than the information of cross-sectional time points. Existing methods exploring viewers' gifting behavior cannot make full use of all the multimodal information available to the viewers, such as the live streaming scene, live sound and live comments. In this research, we try to remedy these defects and consider the multimodal time-series information for gifting prediction.

2.2. Prediction tasks in live streaming

Most of the research on prediction in live streaming context focuses on viewer engagement related metrics, like the number of viewers, viewer comments, and viewer gifting. As for the research on viewer number, Chen et al. (2021) utilized SVM, K-Means and other machine learning methods to predict the number of viewers based on the text information of viewer comments in game live streaming. Similarly, Wehner et al. (2020) used random forest method to predict the viewer number over the course of a soccer match. There are also studies on live comments themselves. For example, the content properties and distribution patterns of live comments have also been extensively studied (He et al., 2017).

Regarding predicting the viewer gifting, Tu et al. (2018) made use of the decision tree model, which treats the number of fans and followings, gender and some other streamers public features as inputs and outputs the number of gifts the streamer has received. Liu et al. (2022) established a structural equation model to predict the gifting number and gifting amount after processing the viewers' real-time comments through traditional machine learning. The number and the length of live shows, as well as the number of viewer comments are used as features to predict which streamers will receive gifts and which viewers will gift those (Jia et al., 2021). They also choose traditional machine learning methods, like logistic regression, random forest, and support vector machine, which are not suitable for multi-modal scenarios. These studies focus on predicting static metrics, while ignore the importance of viewers' real-time gifting prediction.

To sum up, in the live streaming prediction field, the majority of the existing studies first analyze the attributes related to viewer engagement, and then use these attributes as features to make predictions through structural equation modelling, regression models or machine learning methods (like decision tree, random forest). Due to the limitations of these traditional models, the prediction tasks can only utilize some simple numeric features. The selection of these attributes depends heavily on experience, which inevitably leads to the problem of omitted variable bias and the live streaming content information loss. In addition, previous research is based on cross-sectional data and the macro attributes of the entire segment of the live streaming. The ignorance of the effect of time accumulation on viewer gifting makes it unable to predict viewer gifting behavior in real time.

In some live streaming recommendation systems, the real-time problem of donation prediction is solved, but it still depends on historical data. For example, considering the relationship between the streamers and viewers, asymmetric communications, and the trade-off between personal interests and group interactions in live multi-streaming, Lai et al. (2020) proposed a recommendation system called Multi-stream Party Recommender System, which extracts latent features through donation-response tensor factorization for donation and multi-stream party recommendation. Their recommendation system is mainly designed for live multi-streaming, and it requires the information of streamer relations and the viewer social network in advance. However, the live streaming we study is not limited to live multi-streaming, and the correlation among the streamers can be small. We mainly predict the gifting of a single channel through real-time information, such as live streaming video and viewer live comments.

2.3. Multimodal fusion

Multimodal fusion refers to the concept of processing and integrating information from different modalities so that the fused information can be used for downstream prediction tasks, such as classification and regression (Baltrušaitis et al., 2018). This technology is applying widely in various fields, such as in multimodal emotional recognition (Soleymani et al., 2011), and medical image analysis (James & Dasarathy, 2014). Due to the different expression modes of different modalities, the information between the modalities is often overlapping and complementary. In empirical studies, multimodal fusion can often obtain more robust prediction results compared to single mode, which is well verified in the Audio-visual Speech Recognition (AVSR) community (Potamianos et al., 2003). There are mainly three architectures for multimodal fusion: early fusion, late fusion and hybrid fusion (D'mello & Kory, 2015; Zeng et al., 2008). Early and late fusions usually occur at the beginning and end of the learning stage, respectively (Uppal et al., 2022; Zhang et al., 2021). The pros and cons of these two methods are closely related to the correlation between modalities. When the correlation between modes is relatively high, late fusion is better than early fusion; otherwise, early fusion is more suitable (Murphy, 2019). While hybrid fusion combines the advantages of early fusion and late fusion, but also increases the complexity of model structure and training difficulty (Zhang et al., 2021). In this paper, we evaluate four different architecture designs: early fusion, late fusion, hybrid fusion and hybrid fusion with residual connection, and select the strategy with the best performance as our final model architecture.

There are three categories of approaches that can help perform multimodal fusion: kernel-based methods, graphical models, and neural networks (Gönen & Alpaydın, 2011; Ngiam et al., 2011). In recent years, deep neural networks have flourished in data fusion (Song et al., 2021; Yang et al., 2022). Compared with the other two methods, deep neural networks possess the following three advantages: Firstly, deep neural network approaches' big capacity can handle large amount of data (Jan et al., 2019). Secondly, neural architectures make it possible to train both the multimodal representation and the fusion module in an end-to-end manner (Ariav & Cohen, 2019). Thirdly, they often perform better and are more capable of learning complex decision boundaries compared to non-neural network methods (Baltrušaitis et al., 2018). When dealing with multimodal information, deep neural networks still face several challenges, such as redundant inactive parameters and ineffective cross-modal information utilization (Gao et al., 2020). These problems exist even in today's state-of-the-art multimodal emotion recognition model MulT (Tsai et al., 2019), which inspires our

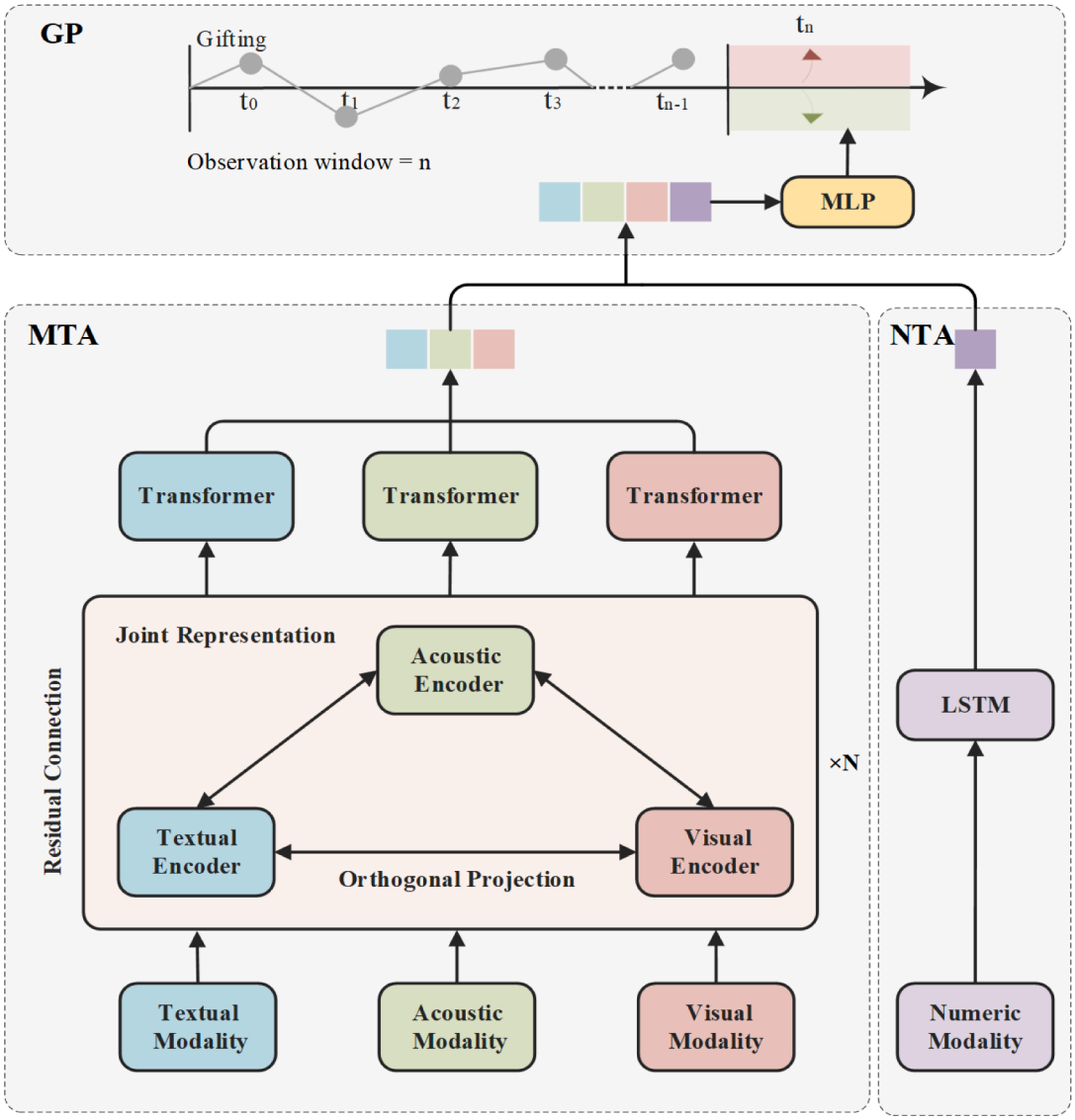


Fig. 1. The overall method of MTM.

model. Therefore, we try to reduce the redundant parts of MulT and improve the effect of cross-modal information fusion. On the premise of ensuring the effect of the method, we hope that the proposed module and fusion strategy can accelerate reasoning, possess high robustness and transferability.

3. The proposed method

In this section, we propose a multimodal time-series method (MTM) for gifting prediction, which consists of three modules including multimodal time-series analysis (MTA) module, numeric time-series analysis (NTA) module and gifting prediction (GP) module, shown in Fig. 1. We will next introduce each of the three modules in detail.

As can be seen from Fig. 1, the first module is the multimodal time-series analysis (MTA) module, which is the core part of our method. From a macro perspective, our MTA can be used alone or as a plug-and-play part of the multimodal information fusion task. First, the raw multimodal inputs (including textual modality, visual modality, acoustic modality) are preprocessed with specialized pretrained models to get the initial representation. Second, we follow the feed-forward fusion process of Tsai et al. (2019) and combine the encoding part with the multimodal fusion part. Each layer of the transformer encoder that process modality A will obtain additional information from modality B and C, which enables information exchange within the same level of representations. To ensure efficient information interaction, we propose the orthogonal projection (OP) model and the joint representation layer. We also add residual connections to the encoder layers to ensure that all low, middle and high-level features can serve the final prediction task properly.

Table 1
Numeric features.

Features	Description
#gifts	the number of gifts
#gift_users	the number of viewers who paid for gifts
#comments	the number of live comments
#comment_users	the number of viewers who sent live comments

Finally, we employ sequence models, such as self-attention transformer (Vaswani et al., 2017), to integrate each modality's serial information.

The second module is the numeric time-series analysis (NTA) module, which is made to encode and aggregate the historical numeric information. First, we collect the numeric information. We include the historical gifting records within the observation window. In addition, we consider some user-visible attributes related to the live streaming video and the streamer, such as #comments, #comment_users. All of the minute-level historical numeric features are elaborated in Table 1. Second, we utilize a single layer of LSTM network, which is good at capturing historical numeric information (Zhang et al., 2021). The last hidden state of the LSTM is used to represent the trend information in the numeric features.

The third module is the gifting prediction (GP) module, which aims at integrating the representations and generating the prediction results. First, we concatenate the fused multimodal representation of MTA with the numerical representation of NTA. Second, we run the combined representation through the fully connect networks. More specifically, we use two stacked structures, with a linear layer followed by a ReLU layer and a dropout layer (Rumelhart et al., 1986; Srivastava et al., 2014), to create our multilayer perceptron (MLP) model. The output of MLP corresponds to our final prediction.

Given the fact that the NTA and GP modules are widely used in previous studies, we will only emphasize on the MTA in the following sections. Section 3.1 covers the data preprocessing part of MTA. Then, the three key designs of the multimodal fusion strategy will be discussed in Section 3.2. For the final part of MTA, we present the serial information integration in Section 3.3.

3.1. Data preprocessing

The raw multimodal data is collected and preprocessed. The details of the multimodal data including textual modality, visual modality, and acoustic modality are introduced as follows.

Textual modality. Live comment, also known as "Danmaku" comments, is a new commentary form which allows comments to be displayed alongside the video. This kind of comment design can promote viewers interaction and closely relates to user engagement (Wu et al., 2019). To utilize the viewer participation information contained in the live comments, we extract the textual features of the live comments as part of the model inputs. Then, we adopt a Chinese version[3] of the state-of-the-art language model called PERT (Cui et al., 2022) to encode the live comments. This model achieves excellent results in several NLP tasks, easy to transfer to other domains (Zhai et al., 2022), and is good at handling informal Chinese dialogue (Pan et al., 2022). Mean pooling, a technique that involves taking the average of sentence embeddings, is commonly used to obtain the overall embedding of a text modality. This helps maintain robustness and avoid overfitting problems in most cases (Lee et al., 2018). Therefore, we perform mean pooling over the representation of each live comment, and get the embedding of the textual modality corresponding to each video segment, whose dimension is 1024-D.

Acoustic modality. Considering that the background music and vocal part might contribute differently to our task, we treat them separately. Firstly, we adopt the Spleeter tool (Hennequin et al., 2020) to separate the background music and voice from the video. Secondly, we use different versions of pretrained wav2vec model (Baevski et al., 2020) to deal with the two parts. This model has strong transfer ability (Wang et al., 2022; Yi et al., 2021) and robust to noise (Zhu et al., 2022). For the vocal part, we choose the wav2vec2.0 model pretrained on Chinese common voice dataset[4]. While for the background music, we employ the version which is pretrained for music genre classification task[5]. After obtaining the two 1024-D vectors, we concatenate them as the final acoustic representation (2048-D).

Visual modality. Consistent with previous research on live streaming (Li et al., 2021; Lin et al., 2021), we treat 1 min live streaming videos as the minimum basic processing unit of our dataset. Then, to simplify the calculation and memory resource, we take a frame every ten seconds to represent the entire one-minute video, and get six images for each unit. After that, we send each image into the pretrained vision transformer (ViT) (Dosovitskiy et al., 2020) and get the last hidden state as the visual embedding. This ViT model also possesses competitive transferability and robustness for new tasks (Bhojanapalli et al., 2021). Similar to the treatment of the textual modality, we refer to the approach used in past research (Jiang et al., 2019) and adopt the mean pooling method for all the frames to obtain the final visual representation of the one-minute video, which is a 1024-D vector.

[3] https://github.com/ymcui/PERT/blob/main/README_EN.md
[4] https://huggingface.co/ydshieh/wav2vec2-large-xlsr-53-chinese-zh-cn-gpt
[5] https://huggingface.co/m3hrdadfi/wav2vec2-base-100k-gtzan-music-genres

3.2. Multimodal fusion strategy

In this section, we elaborate three key designs of the multimodal fusion strategy of our MTA module. Roughly speaking, our MTA module consists of three branches of transformer encoders, where each branch deals with one kind of modal information. In the front part of each branch, we adopt N transformer encoders similar to the cross-attention design to promote multimodal information interaction (Vaswani et al., 2017). In particular, we not only consider the method for inter-modal information interaction at the same level of encoders (Sections 3.2.1 and 3.2.2), but also better retention of intra-model information of different levels (Section 3.2.3). In the back part of each branch, we use the self-attention transformer encoders to aggregate sequence information.

3.2.1. Orthogonal projection

We consider textual modality (T), acoustic modality (A), and visual modality (V), with three encoded sequences from the i-th encoder X_T^i, X_A^i and X_V^i. In the past many studies, bilinear models show strong multimodal fusion and mixing ability (Yu et al., 2017; Kim et al., 2017; Ben-Younes et al., 2019). The common feature of these methods is to find the relevant parts among the modes and form joint representations by means of matrix decomposition. All modes are encoded and respectively then converged to form a unified representation through bilinear fusion modules. Then, such joint representation vector that extracts common features from each modality is used for downstream tasks. Other researchers, such as Xiao et al. (2022), have begun to notice the complementation and substitution effects between different modalities, and employ different methods (such as attention mechanisms and loss functions) to improve the interaction of modality-specific features. These studies have inspired us to solve this problem from a different point of view. We intend to retrieve irrelevant information from other modes as a supplement of the target mode. Instead of finding the corresponding part between modes, we try to find the orthogonal parts from other modes to enhance the target mode. At the same time, we are also inspired by the crossmodal attention mechanism in MulT (Tsai et al., 2019) that feed the low-level information into the later layers of encoder, which repeatedly forces the low-level information from other modalities to participate in the each-level attention calculation with the target modality. Their design pays too much attention to the low-level features, which may make the model less capable of learning high-level information. We propose a better way to utilize crossmodal information efficiently, i.e., from modalities A, V to T. Different from using two separate transformers to respectively take in the low-level representation from other two modalities like MulT, we only employ one transformer encoder to reinforce the target modality by utilizing the fused same-level information from the other two modalities. We change the traditional routine that key and value are derived from single mode to the fused mode (simple swaps and linear transformations). To fuse the information from the other two modalities, we propose the orthogonal projection (OP) model. The motivation for OP model is to preserve the parts of other modalities that are orthogonal to the target modality and remove duplicate information to prevent redundancy. Then, the fused external modal information can be involved in the attention calculation process of the target modality. An illustration demonstrating the application of the OP model in a live streaming context can be found in Appendix A.

Our orthogonal projection model can be described as the following steps (here we set T as the target modality):

First, we conduct L2 normalization (Perronnin et al., 2010) over each encoded sequence X_T^i, X_A^i and X_V^i, and get $X_T^{i'}$, $X_A^{i'}$ and $X_V^{i'}$. We then measure the correlation between the corresponding positions of the two modal vectors by calculating the dot product as (1).

$$\mathrm{Corr}_{TA}^i = X_T^{i'} \cdot X_A^{i'}, \ \mathrm{Corr}_{TV}^i = X_T^{i'} \cdot X_V^{i'} \tag{1}$$

Second, we use SoftMax operation to approximate the magnitude of the correlation between mode T and the corresponding position of mode A (or V) with each element of Corr_{TA} (or Corr_{TV}).

$$\mathrm{Corr}_{TA}^{i'} = SoftMax\left(\mathrm{Corr}_{TA}^i\right), \mathrm{Corr}_{TV}^{i'} = SoftMax\left(\mathrm{Corr}_{TV}^i\right) \tag{2}$$

Afterwards, we perform a 1-x operation (for each input x, a 1-x is returned, each position is subtracted by (1) on the correlation value and get the dissimilarity vector, which measures the level of difference between the corresponding positions of the two modes' representation. Then, we multiply the original representations (X_A^i and X_V^i) with these dissimilarity vectors and obtain the orthogonal parts of the information to the target modality. These parts eliminate the information duplication between other modes and the target mode, so as to retain the part orthogonal to the target mode vector for next step.

$$\mathrm{Orth}_{TA}^i = X_A^i \cdot \left(1 - \mathrm{Corr}_T A^{i'}\right), \ \mathrm{Orth}_{TV}^i = X_V^i \cdot \left(1 - \mathrm{Corr}_{TV}^{i'}\right) \tag{3}$$

Third, we add these orthogonal parts to the target modality, and get the fused latent adaption of modality A, V to T. Here we retain the original representation of mode T and add the supplementary information from other modes.

$$Y_T^i = \mathrm{OP}\left(X_T^i, \ X_A^i, X_V^i\right) = X_T^i + \mathrm{Orth}_{TA}^i + \mathrm{Orth}_{TV}^i \tag{4}$$

After that, the Y_T^i can participate in the cross-attention process of the i-th encoder of modality T (as the source of key and value).

3.2.2. Joint representation

We employ one layer of the transformer encoder with cross attention (Vaswani et al., 2017) as our basic block (Basic_B), and change the information exchange mechanism of crossmodal attention block in MulT (Tsai et al., 2019). For each mode in MulT, there are two paths of encoders that respectively receive the local mode and the low-level signals (Key/Value pairs) from the other mode. While in our MTM, information interactions happen in the same level of cross encoders. For each modality, there are N cross attention layers, we

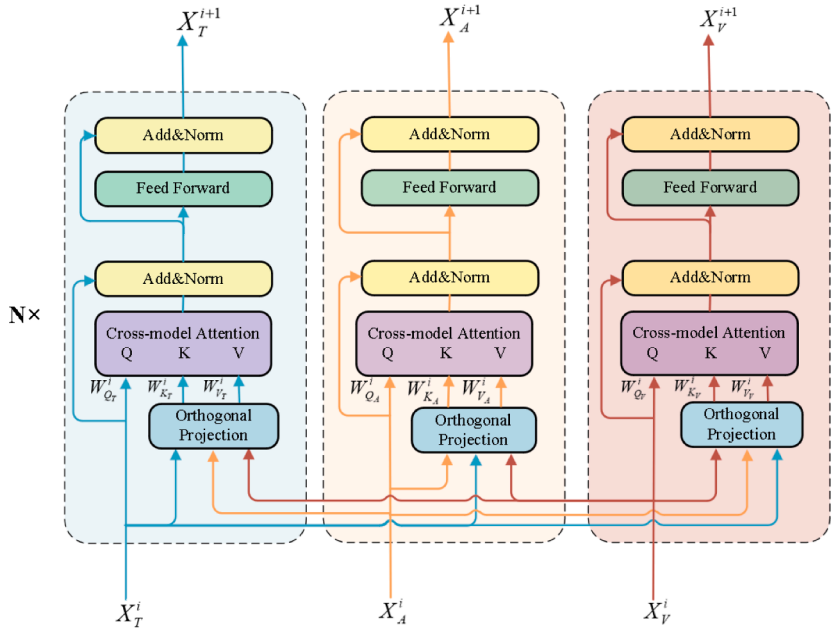


Fig. 2. Demonstration of the joint representation layer.

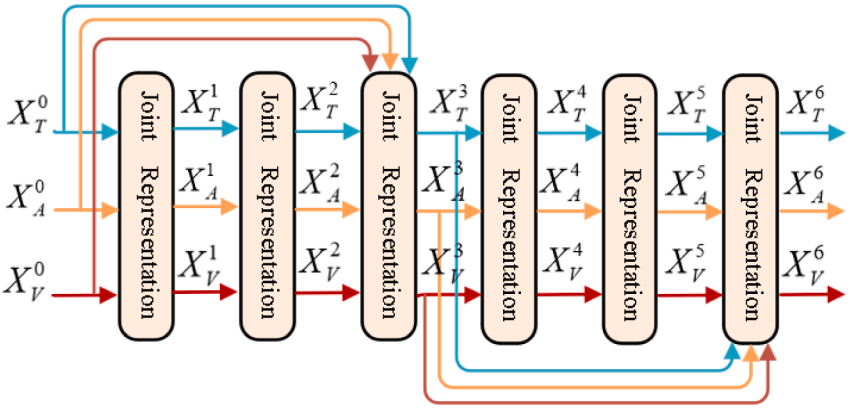


Fig. 3. Residual connections within joint representation layers ($N = 6$).

call the same level of all the modalities' layers as a joint representation layer (see Fig. 2). Within each joint representation layer, there are three basic encoders, each of which deals with one modality. The basic encoder receives information from the current mode and the fused information from the other two modes with the help of OP model (see Section 3.2.1). Thus, the Query inherits the current mode's information, while the orthogonal projection model helps complements information from other modes, forming the Key and Value. This joint representation layer facilitates multimodal information interaction at the same level, and can be stacked multiple times to get more levels of representations.

To deal with the inputs from three modalities (visual, acoustic, and textual modality), we use 1D temporal convolution layers to capture sequence information and prepare their dimensions for the following encoders. Then, the representations go through N (=6) stacked joint representation layers, gaining additional information from other modalities. The forward process of our joint representation layers (i = 0, 1…, N-1) can be described as follows:

$$X_T^{i+1} = Basic\_B\left(X_T^i W_{Q_T}^i, Y_T^i W_{K_T}^i, Y_T^i W_{V_T}^i\right), Y_T^i = OPM\left(X_T^i,\ X_A^i, X_V^i\right) \tag{5}$$

$$X_A^{i+1} = Basic\_B\left(X_A^i W_{Q_A}^i, Y_A^i W_{K_A}^i, Y_A^i W_{V_A}^i\right), Y_A^i = OPM\left(X_A^i, X_T^i,\ X_V^i\right) \tag{6}$$

$$X_V^{i+1} = Basic\_B\left(X_V^i W_{Q_V}^i, Y_V^i W_{K_V}^i, Y_V^i W_{V_V}^i\right), Y_V^i = OPM\left(X_V^i,\ X_T^i, X_A^i\right) \tag{7}$$

Table 2
Dataset split result.

Dataset	Total	Train	Test	Valid
Our dataset (window=5)	40375	28262	8075	4038
Our dataset (window=10)	44185	30930	8837	4418
CMU-MOSI	2199	1284	686	229
CMU-MOSEI	22777	16265	4643	1869
IEMOCAP	4453	2717	938	798

where $W^i_{Q_m}, W^i_{K_m}, W^i_{V_m}(\mathrm{m} \in (T,A,V))$ refer to weights, and the inputs of Basic_B correspond to Query, Key, and Value of cross attention in turn.

3.2.3. Residual connection

Considering the fact that each of the joint representation layers only enables information interaction at the same level, inspired by He et al. (2016), we add residual connections to the Basic_B to ensure the preservation of all levels of information (see Fig. 3). Specifically, the raw input (X^0_T, X^0_A, X^0_V) is added to the output of the 3rd joint representation layer, which together make the input (X^3_T, X^3_A, X^3_V) of the 4th joint representation layer. This residual connection forces the model of each path to reinforce the information from the respective modes. While ensuring the same level of information interaction, our residual connection can preserve both low-level and high-level information.

3.3. Serial information integration

In order to better aggregate the temporal information after the previous multi-modal fusion modules, we add the serial information integration part. Afterwards, following the MulT settings (Tsai et al., 2019), the fused representation of each modality (X^N_T, X^N_A, X^N_V) has to go through a sequence model. Here we choose the self-attention transformer encoders, to capture the temporal information (Vaswani et al., 2017).

Specifically, we input the hidden states of the three modes from the last joint representation layer into the three self-attention encoders respectively, thus obtaining the final representation of each modality containing multimodal information and self-attention information. So far, this is the overall structure of our proposed MTA model, which is the core module of MTM.

4. Empirical analysis

4.1. Data description and evaluation metrics

Besides testing our model on our dataset, we further evaluate the MTA on three standard multimodal datasets to prove the model's improvement in different scenarios. The result of dataset division is shown in Table 2. To meet the real-time requirements of live streaming scene, the trade-offs between performance/speed should also be taken into consideration (Zhang et al., 2021). For each of the following tasks, we also add evaluation metrics for the model structure and time complexity (Abdu et al., 2021): the number of parameters (#parameters), floating-point operations per second (FLOPs) and average training time per epoch in seconds (time/epoch). These indicators can be used to measure the performance of the real-time prediction models. Under consistent performance, it is better for a model to have fewer parameters, smaller computational volume, and shorter training time. We will elaborate on the datasets below.

Our dataset. Our dataset originates from Bilibili (https://live.bilibili.com/), a popular live streaming video platform, which attracts nearly one of every two young people (under 24) in China[6] (Ding et al., 2022). All the data is open and freely accessible on the Internet. We build the dataset from two sources: the live streaming video and the gifting records. Specifically, we track 129 long-term active streamers who have more than 1 year of live streaming experience and considerable fan volumes among the 7 segments (PC games, entertainment, learning, mobile games, life, radio, online games). Our dataset covers the period from September 15, 2021 to October 4, 2021. All videos were collected from the active streams of the selected streamers during the whole day. We obtain live streaming videos of over 1809.69 h and 358002 gifting records and the video clips are stored on a high-capacity server at a frequency of 60 frames per second and a resolution of 1920*1080. After aligning the video content and the gifting records, we transform the data into a time-series format with the fixed-length sliding time window method.

For robustness consideration, the observation time window of the one-min gifting prediction task is set at 5 and 10 mins respectively. To avoid the endogenous difference in the amount of gifting caused by the unobservable attributes of the streamers themselves, we binarize the prediction target according to whether it is greater than the mean of gifting in the observation window. After removing vacant values and samples with very low gifting amount, we obtain two datasets with sample sizes of 40375 and 44185, corresponding to observation time window of 5 and 10, respectively. We randomly split the data set for training, testing, and validation in a 7:2:1

[6] Bilibili. (2021). Industry report. https://www.bilibili.com/read/cv9161223.

Table 3
Hyperparameters for the tasks.

Hyperparameters	Our dataset (window = 5)	Our dataset (window = 10)	CMU-MOSI	CMU-MOSEI	IEMOCAP
Loss Function	Focal Loss	Focal Loss	L1Loss	L1Loss	CrossEntropyLoss
Initial Learning Rate	1e-3	1e-3	1e-3	1e-3	2e-3
Batch Size	1024	1024	1024	1024	1024
# of Crossmodal Blocks	6	6	6	6	6
# of Crossmodal Attention Heads	12	12	10	10	10
# of Epochs	40	40	100	40	60
Temporal Convolution Kernel Size	1	1	1	1	3
Transformers Hidden Unit Size d	60	60	40	40	40
Focal-alpha	0.7	0.75			
Focal-gamma	0.9	1.1			
Parallel training or not	No	No	Yes	Yes	Yes

Where the "# of Crossmodal Blocks" refers to how many layers of encoder in total.

ratio, and the ratio of positive and negative samples is about 1:3. Because our task is an unbalanced binary classification problem, we choose precision, recall and F1 score as the evaluation metrics.

CMU-MOSI & CMU-MOSEI. CMU-MOSI (Zadeh et al., 2016), a collection of 2199 monologue video clips, which includes pre-processed acoustic, visual and textual features. There are 98 speakers and the total length of CMU-MOSI video is about 2.5 h. Meanwhile, the size of CMU-MOSEI (Zadeh et al., 2018) is about 10 times that of CMU-MOSI. It is made up of 23453 video segments, 1000 speakers, whose total length of video reaches 65 h. These two datasets are often used to evaluate models' multimodal fusion capabilities (Wang et al., 2019).

Both of CMU-MOSI and CMU-MOSEI are manually labeled with 7 sentiment scores ranging from -3 (strongly negative) to 3 (strongly positive). We utilize the practice of prior works (Tsai et al., 2019; Yang et al., 2020) and choose the following evaluation metrics: 7-class accuracy, binary accuracy, F1 score, mean absolute error (MAE), and the correlation value of the model's prediction with human.

IEMOCAP. IEMOCAP (Busso et al., 2008) consists of 10000 video clips recorded by 10 speakers. The 12 h of audio-visual data contains video, voice, facial motion capture, text transcription. Each segment is annotated for the presence of 9 emotions as well as valence, arousal and dominance. We follow the practice of Wang et al. (2019) and Tsai et al. (2019) and choose 4 of the emotions (happy, sad, angry and neutral) for emotion recognition task. Due to the fact that it is a multilabel task, we report the binary classification accuracy and F1 score for each emotion type.

4.2. Experimental settings

We implement our MTM on our dataset in PyTorch on single NVIDIA RTX3090 GPU with 24GB memory. The batch size of our dataset is set as 1024. To alleviate the imbalance of positive and negative samples, we choose focal loss (Lin et al., 2017) as the loss function and the hyperparameters setting details are attached in Appendix B. We adapt Adam optimizer with an initial learning rate of 1e-3, and the grid search results of learning rates can be found in Appendix C. Besides, we decay the learning rate in half every 20 epochs. In the network, the embeddings of each modality go through 1D temporal convolutional layers to maintain neighborhood information and covert to fixed dimension, before they are sent into the encoders. The dimensions of the model's embeddings layer (see Appendix D) and the number of crossmodal blocks (see Appendix E) are also determined via grid search. For all the tasks, we set the number of layers of transformer encoder to 6. Besides, the determination of other hyperparameters is also based on the grid search method. Our settings on the standard datasets are similar to those of MulT (Tsai et al., 2019). The details of hyperparameters are shown in Table 3.

In order to better evaluate the model effects, we adopt a variety of multi-modal fusion methods as our baseline models. The main difference between these baseline models and our model lies in the multi-modal fusion part, while the necessary encoding layer and sequence model (self-attention transformers) remain consistent for a fair comparison. Besides, inspired by Zhang et al. (2021), we separately test the effect of early fusion, late fusion, hybrid fusion and hybrid residual fusion with designs on our model.

All of the competing methods are shown as follows:

- MFB (Yu et al., 2017): Multi-modal Factorized Bilinear (MFB) pooling approach utilizes a bilinear interaction to efficiently and effectively combine multi-modal features by constraining the rank of each 3rd mode slice matrix in the tensor.
- MLB (Kim et al., 2017): Multimodal Low-rank Bilinear attention networks (MLB) is a low-rank bilinear pooling method that uses the Hadamard product to provide an efficient attention mechanism for multimodal learning. This method involves a bilinear interaction where the tensor is expressed as a CP decomposition.
- Mutan (Ben-Younes et al., 2017): Multimodal Tensor-based Tucker decomposition (Mutan) generates a bilinear interaction where the tensor is expressed as a tucker decomposition, and the core tensor shares the same low-rank constraint with MFB.
- MFH (Yu et al., 2018): Multimodal Factorized High-order pooling (MFH) method is a higher order fusion composed of cascaded MFB.

Table 4
Results on our dataset when observation window is set to 5 min.

Methods	#Parameters	Training Time (sec/epoch)	FLOPs	Precision (%)	Recall (%)	F1(%)
MFH	1899782	9.533	4311360	65.94	53.07	56.58
MFB	1534982	8.260	1534982	65.80	25.54	13.91
MLB	1388582	9.190	3807360	61.21	27.86	20.65
Mutan	78309382	15.692	80679360	62.25	32.88	31.03
Block	5349382	9.773	7719360	70.71	23.98	9.79
Joint-encoding	8235302	10.153	39909060	62.48	31.15	27.54
MulT	5498522	11.872	25514160	65.57	27.69	19.11
EF-MTM w/o res	2222882	8.782	9433500	68.12	25.37	13.29
LF-MTM w/o res	2222882	8.854	9433500	65.80	30.59	25.31
MTM w/o res	2277782	10.818	9748500	64.54	64.73	64.63
MTM	2277782	10.107	9748500	64.10	65.11	64.59

Table 5
Results on our dataset when observation window is set to 10 min.

Methods	#Parameters	Training Time (sec/epoch)	FLOPs	Precision (%)	Recall (%)	F1(%)
MFH	1899782	10.636	7895160	69.19	36.64	39.85
MFB	1534982	9.203	1534982	69.00	41.82	46.62
MLB	1388582	9.166	7391160	68.98	28.05	26.05
Mutan	78309382	10.198	84263160	69.56	47.72	53.05
Block	5349382	10.098	11303160	70.17	37.14	40.26
Joint-encoding	8235302	10.426	79954560	69.51	35.44	38.04
MulT	5498522	17.879	50984160	68.75	27.28	24.66
EF-MTM w/o res	2222882	14.099	18888000	69.93	51.43	56.64
LF-MTM w/o res	2222882	13.697	18888000	68.87	65.51	67.05
MTM w/o res	2277782	16.639	19518000	68.80	68.35	68.57
MTM	2277782	16.204	19518000	68.79	70.45	69.59

- Block (Ben-Younes et al., 2019): Block is an advanced framework for multimodal representation utilizing block-term tensor decomposition and the concept of block-term ranks to strike the optimal balance between expressiveness and complexity in fusion models. It enables the capture of subtle intermodal interactions while preserving powerful mono-modal representations.
- Joint-encoding (Delbrouck et al., 2020): Joint-encoding is the state-of-the-art model on CMU-MOSEI dataset, which is a useful transformer-based joint encoding method with modular co-attention and a glimpse layer targeting at multimodal emotion recognition. For fair comparison, we replace the joint representation layers of our method with Joint-encoding's co-attention and glimpse layer design. We then modified the code provided by the authors to accommodate all our 5 datasets.
- MulT (Tsai et al., 2019): Multimodal Transformer (MulT) utilizes the directional pairwise cross-modal attention to enable information interaction between modalities, and achieve the state-of-the-art performance on multimodal time-series prediction tasks.
- EF-MTM w/o res (Early fusion MTA without residual connection): Different from our MTM, we replace each basic cross attention block of the i-th (i=1,2, …, N-1) joint representation layer with self-attention block in MTA. Only the first ($i = 0$) joint representation layer enables cross-modal information interaction.
- LF-MTM w/o res (Late fusion MTA without residual connection): Similar to early fusion strategy, we only keep the cross-attention blocks in the last (i=N-1) joint representation layer.
- MTM w/o res (Hybrid fusion MTA without residual connection): We remove the residual connections within the multimodal cross-modal blocks of each modality, only keeping the feed-forward fusion process.
- MTM (Our proposed multimodal time-series method): We not only use N layers of joint representation, but also employ residual connections to ensure multi-scale information fusion.

4.3. Experimental results and discussions

4.3.1. Main results

To verity the effectiveness our MTM, we train and test all the models we mentioned in Section 4.2 on our dataset, including MFH, MFB, MLB, Mutan, Block, Joint-encoding, MulT, EF-MTM w/o res (Early fusion MTA without residual connection), LF-MTM w/o res (Late fusion MTA without residual connection), MTM w/o res (Hybrid fusion MTA without residual connection), and MTM (Our proposed multimodal time-series method). For the consideration of robustness, we set the observation window of our dataset to 5 and 10 mins respectively. The main results on our datasets (window = 5, 10) are shown in Tables 4 and 5.

As can be seen from Table 4, our MTM and MTM w/o res model can achieve higher F1 score compared to other baseline models. Our proposed method requires less than half the number of parameters and has lower computational complexity than models with similar structure (e.g., MulT and Joint-encoding), and requires less training time. Compared with other baseline models with fewer parameters, our MTM can return most of the relevant results with higher recall. Meanwhile, by comparing the performance of early, late and

Table 6
Results of comparisons between our orthogonal projection model and other models.

Methods	#Parameters	Training Time (sec/epoch)	FLOPs	Precision (%)	Recall (%)	F1(%)
Observation window = 5 mins						
MTM-ConcatMLP	8350982	12.403	40024500	64.25	71.67	66.70
MTM-LinearSum	6144182	11.296	28864500	62.68	25.21	13.44
MTM	2277782	10.107	9748500	64.10	65.11	64.59
Observation window = 10 mins						
MTM-ConcatMLP	8350982	16.024	80070000	71.13	32.73	33.42
MTM-LinearSum	6144182	14.247	57750000	69.28	29.44	28.48
MTM	2277782	12.204	19518000	68.79	70.45	69.59

Table 7
Results of modal ablation experiments of MTM on our datasets.

Methods	#Parameters	Training Time (sec/epoch)	FLOPs	Precision (%)	Recall (%)	F1(%)
Observation window = 5 mins						
MTM w/o T	1556642	7.172	6575880	64.41	53.42	56.86
MTM w/o A	1510562	6.279	6345480	62.82	27.91	20.30
MTM w/o V	1556642	6.953	6575880	65.81	45.99	49.18
MTM	2277782	10.107	9748500	64.10	65.11	64.59
Observation window = 10 mins						
MTM w/o T	1556642	11.789	13165680	68.92	53.94	58.86
MTM w/o A	1510562	11.179	12704880	70.17	23.54	16.72
MTM w/o V	1556642	11.262	13165680	68.36	28.94	27.85
MTM	2277782	12.204	19518000	68.79	70.45	69.59

hybrid fusion methods, we find that hybrid fusion can better obtain multimodal information under the circumstance of short observation window. The reason for this phenomenon may be that multi-scale signals are more needed when there is less available sequence information. More information interaction between modes can alleviate the problem of the lack of useful information caused by the short observation time window.

In Table 5, our MTM stands out with high F1 score. It can be inferred from these two tables that the performance of the models will improve with the lengthening of the observation window. Although the performance of other baseline models is slightly improved with the time window becoming longer, their recall values are still at a low level. Besides, MulT cannot effectively deal with longer sequence information, resulting in little improvement. One possible explanation is that MulT forces every level of the crossmodal attention block to take in the low-level signals form the source modality, which brings about information redundancy and breaks the balance between high-level and low-level information. Our MTM model, on the other hand, is better at capturing all levels of information and achieves the best overall performance.

4.3.2. Comparisons of different fusion models

When applying hybrid fusion strategy, the complexity of the fusion model matters. Thus, we compare our OP model with simple models with additional parameters, such as Concat-MLP and Linear-Sum method (Ben-Younes et al., 2019). Table 6 illustrates our model's performance using different fusion models.

Thus, when using a fusion model with additional parameters, the number of parameters will increase dramatically, which may cause overfitting. For each joint representation layer (with 3 modalities), 3 fusion blocks are needed. While for the whole model (with N representation layers), there will be 3*N fusion blocks. Therefore, our proposed lightweight OP can meet the requirements of deeper networks and more modalities, and can be easily deployed to mobile devices. In the case of real-time prediction, faster inference speed can also be achieved to ensure the timeliness of results. In addition, we can see that the recall of our MTM with ConcatMLP method drop 38.94% when the time window switch to 10 mins. A larger number of model parameters can cause a decrease in recall because the model may become too complex and learn the noise and randomness in the training data instead of the true patterns and regularities (Hornik, 1991), which can lead to overfitting and cause the model to perform poorly on longer time-series data.

4.3.3. Ablation study on modalities

To justify our use of the information from three modalities, we performed the following ablation experiments on the mode types. Here, we introduce three models for comparison: MTM w/o T (MTM without textual modality), MTM w/o A (MTM without acoustic modality), MTM w/o V (MTM without visual modality).

As can been seen from Table 7, using all the three modalities does improve MTM's performance. Removing the acoustic information can bring about the most damage to the recall, visual information the next. In addition, the removal of textual information has relatively little impact compared to the other two modalities. To conclude, using all three modes leads to optimal results on our datasets.

Table 8
Results of MTA on the standard CMU-MOSI datasets.

Metric	MAE	Corr	Acc_7(%)	Acc_5(%)	Acc_2(%)	F1(%)
MFH	1.1358	0.5324	28.13	32.36	70.12	69.94
MFB	1.1183	0.5704	28.43	31.63	71.65	71.47
MLB	1.1401	0.5356	28.13	32.51	70.88	70.75
Mutan	1.1228	0.5463	28.13	30.90	72.87	72.71
Block	1.1045	0.5760	29.15	32.94	73.78	73.93
Joint-encoding	1.2666	0.4626	23.91	27.70	64.02	63.84
MulT	1.0296	0.6591	30.32	37.46	75.91	75.77
EF-MTA w/o res	1.3849	0.3095	19.10	19.39	55.03	57.57
LF-MTA w/o res	1.5288	-0.0728	16.47	16.47	43.90	52.42
MTA w/o res	1.3476	0.3414	21.28	22.30	61.28	61.67
MTA	1.1602	0.6316	26.97	38.63	78.35	78.24

Table 9
Results of MTA on the standard CMU-MOSEI datasets.

Metric	MAE	Corr	Acc_7(%)	Acc_5(%)	Acc_2(%)	F1(%)
MFH	0.6385	0.6283	48.50	49.77	79.97	79.88
MFB	0.6379	0.6215	48.83	50.16	76.63	76.27
MLB	0.6357	0.6228	48.93	50.29	78.18	78.61
Mutan	0.6524	0.6201	46.69	47.56	78.81	78.64
Block	0.6336	0.6273	48.24	49.49	79.20	79.80
Joint-encoding	0.7046	0.5605	44.07	44.80	77.57	77.56
MulT	0.6331	0.6608	48.87	50.61	80.50	80.37
EF-MTA w/o res	0.8151	0.2301	41.89	41.89	64.39	66.98
LF-MTA w/o res	1.0496	0.0444	27.63	27.63	43.37	47.65
MTA w/o res	0.6524	0.6357	47.56	49.08	80.22	80.30
MTA	0.6483	0.6455	48.20	49.82	80.66	81.00

Table 10
Results of MTA on the standard IEMOCAP dataset.

IEMOCAP	Neutral		Happy		Sad		Angry	
Metric	F1	Acc	F1	Acc	F1	Acc	F1	Acc
MFH	65.62	67.06	82.00	84.97	82.55	81.88	86.91	86.78
MFB	65.71	66.74	81.52	84.75	80.63	81.13	84.82	84.86
MLB	66.97	67.70	82.01	84.01	82.52	82.41	84.99	85.39
Mutan	65.15	65.46	81.98	85.61	75.78	79.53	81.90	83.37
Block	64.75	66.74	81.96	83.69	82.17	81.24	87.94	88.27
Joint-encoding	65.19	65.78	83.51	86.35	81.45	80.81	84.95	84.86
MulT	68.31	68.76	84.33	86.25	83.37	82.62	86.47	86.46
EF-MTA w/o res	67.46	67.48	85.15	86.99	82.46	81.34	87.03	87.10
LF-MTA w/o res	46.46	48.40	72.35	69.94	71.02	73.13	67.20	66.42
MTA w/o res	68.07	68.87	84.25	86.35	82.20	81.13	86.52	86.14
MTA	68.07	69.08	84.47	85.93	82.78	81.88	87.19	86.89

Table 11
Results of structure related metrics on the standard CMU-MOSI dataset.

Metric	#Parameters	Training Time (sec/epoch)	FLOPs
MFH	887161	0.607	15760680
MFB	642361	0.613	15520680
MLB	543961	0.601	15424680
Mutan	77440761	0.689	92272680
Block	4480761	0.706	19312680
Joint-encoding	6293721	0.942	304016040
MulT	2240921	1.395	121125440
EF-MTA w/o res	818081	1.077	43668920
LF-MTA w/o res	818081	1.083	43668920
MTA w/o res	842681	1.257	45168920
MTA	842681	1.275	45168920

Table 12
Results of structure related metrics on the standard CMU-MOSEI dataset.

Metric	#Parameters	Training Time (sec/epoch)	FLOPs
MFH	897241	5.129	16158120
MFB	652441	4.860	15918120
MLB	554041	5.006	15822120
Mutan	77450841	5.652	92670120
Block	4490841	5.614	19710120
Joint-encoding	6303801	10.180	304499880
MulT	2244281	11.187	121293440
EF-MTA w/o res	821441	8.150	43836920
LF-MTA w/o res	821441	8.166	43836920
MTA w/o res	846041	9.320	45336920
MTA	846041	9.189	45336920

Table 13
Results of structure related metrics on the standard IEMOCAP dataset.

Metric	#Parameters	Training Time (sec/epoch)	FLOPs
MFH	900328	0.900	5843200
MFB	655528	0.905	5603200
MLB	557128	0.912	5507200
Mutan	77453928	0.941	82355200
Block	4493928	1.037	9395200
Joint-encoding	6303801	9.572	304499880
MulT	2278688	1.675	41012400
EF-MTA w/o res	855008	1.349	14730720
LF-MTA w/o res	855008	1.345	14730720
MTA w/o res	879608	1.605	15270720
MTA	879608	1.640	15270720

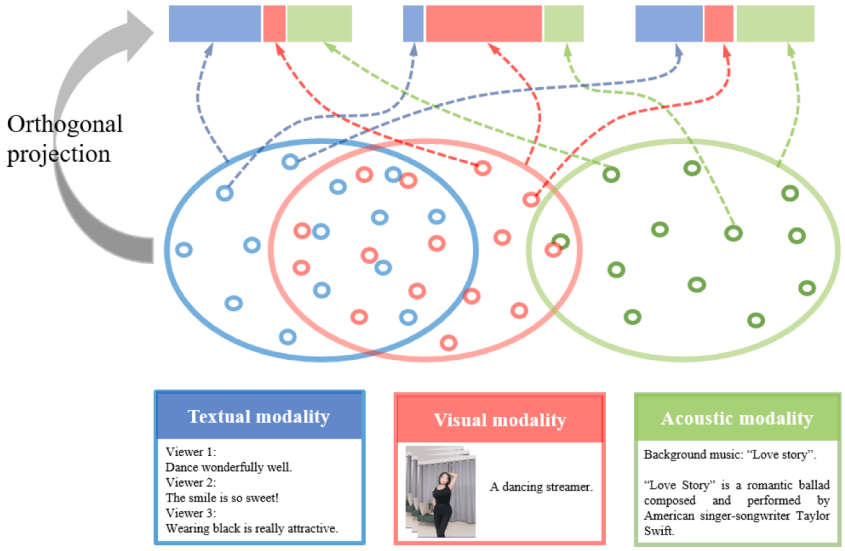


Fig. A.1. An example of the OP model's role in high and low redundancy information.

4.3.4. Robustness check

In order to prove the robustness and transferability of our proposed method, we also validate it on three public standard datasets of different tasks. We separately take out the MTA module in our MTM, which can adapt to the input and output requirements of the standard datasets. Then, we retrain all the models' twenty times on each dataset and report the average performance. The results of emotion recognition task on CMU-MOSI & CMU-MOSEI datasets can be found in Tables 8 and 9.

We follow the evaluation metrics used by Ben-Younes et al. (2019): mean absolute error (MAE), correlation of the model's predictions with human labels (Corr), 7-class accuracy (Acc_7), 5-class accuracy (Acc_5), binary accuracy (Acc_2) and F1 score. We also verity our MTA module on the IEMOCAP dataset of multilabel emotion recognition task, and the results are shown in Table 10. Here, we

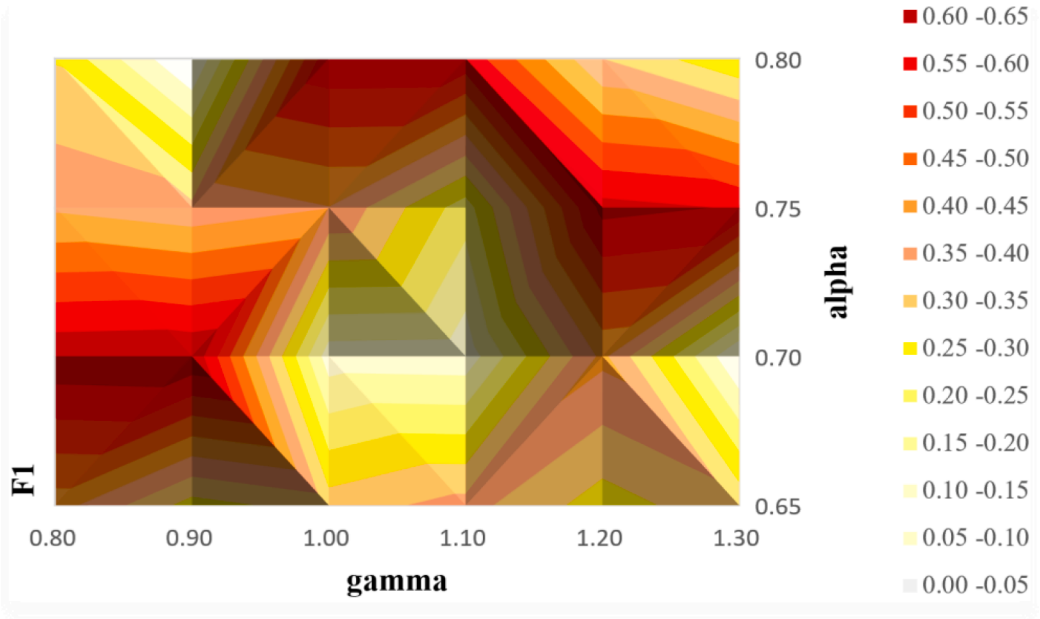


Fig. B.1. Grid search results of focal loss parameters on our dataset(window=5).

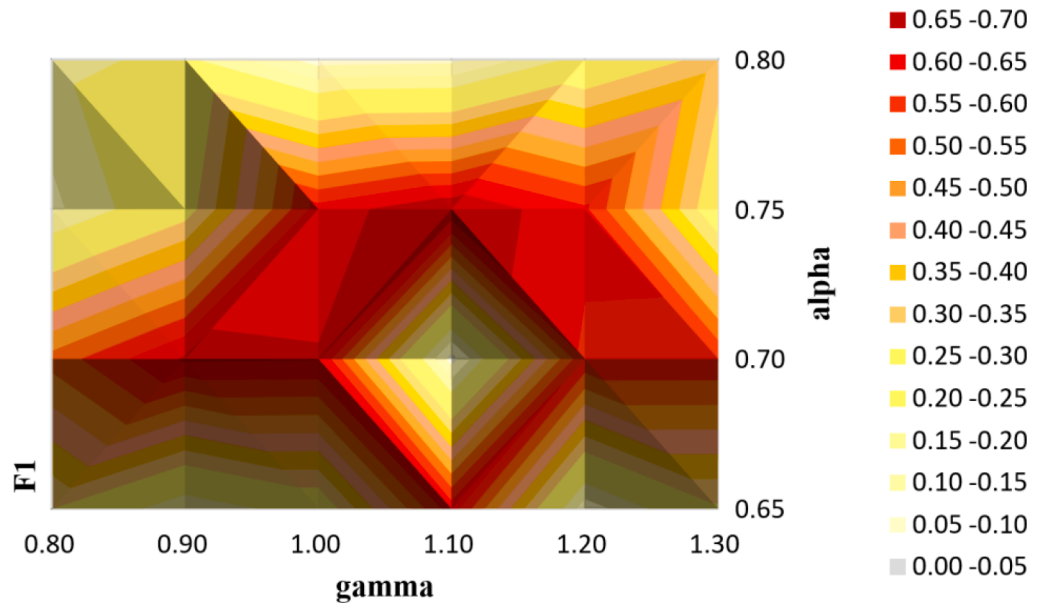


Fig. B.2. Grid search results of focal loss parameters on our dataset (window=10).

report the binary classification accuracy and F1 score of the four emotions (neutral, happy, sad and angry) as suggested by Wang et al. (2019).

As can be seen from Tables 8–10 above, our model performs consistently on the standard dataset and outperforms other baseline models. This provides evidence of the transferability of our model. Then, results about the model structure of MTA module on three standard datasets are shown in Tables 11–13.

As can be seen from Tables 11–13 above, compared to the model MulT, which ranked second in overall performance, our MTA has only 38% of MulT's number of parameters, approximately 37% of its computational cost, and a training time that is approximately 90% as long as MulT's. Although our model does not perform optimally on these three metrics, it does stand out in terms of the stability of its performance. Overall, our model's performance on structure related metrics when transferred to other tasks and datasets remains acceptable.

Table B.1
Grid search results of focal loss parameters on our dataset (window=5).

alpha	gamma	Precision (%)	Recall (%)	F1(%)	alpha	gamma	Precision (%)	Recall (%)	F1(%)
0.65	0.8	0.6350	0.4028	0.4251	0.75	0.8	0.6346	0.3792	0.3909
0.65	0.9	0.6715	0.3215	0.2798	0.75	0.9	0.6321	0.3612	0.3635
0.65	1.0	0.6282	0.3376	0.3243	0.75	1.0	0.6362	0.3846	0.3986
0.65	1.1	0.6212	0.3622	0.3692	0.75	1.1	0.6505	0.2909	0.2237
0.65	1.2	0.6318	0.3127	0.2753	0.75	1.2	0.6542	0.5932	0.6168
0.65	1.3	0.6514	0.3495	0.3368	0.75	1.3	0.6364	0.5700	0.5958
0.70	0.8	0.6293	0.6536	0.6405	0.80	0.8	0.6156	0.3200	0.2961
0.70	**0.9**	**0.6410**	**0.6511**	**0.6459**	0.80	0.9	0.6667	0.2375	0.0926
0.70	1.0	0.6659	0.2489	0.1222	0.80	1.0	0.6310	0.5600	0.5873
0.70	1.1	0.6428	0.2466	0.1177	0.80	1.1	0.6337	0.5560	0.5849
0.70	1.2	0.6284	0.3932	0.4137	0.80	1.2	0.6191	0.3501	0.3499
0.70	1.3	0.6336	0.2391	0.0975	0.80	1.3	0.6505	0.3058	0.2550

Table B.2
Grid search results of focal loss parameters on our dataset (window=10).

alpha	gamma	Precision (%)	Recall (%)	F1(%)	alpha	gamma	Precision (%)	Recall (%)	F1(%)
0.65	0.8	0.6928	0.3037	0.3010	0.75	0.8	0.7009	0.2217	0.1372
0.65	0.9	0.6759	0.2270	0.1534	0.75	0.9	0.6881	0.2847	0.2687
0.65	1.0	0.6773	0.2690	0.2422	0.75	1.0	0.6876	0.5810	0.6213
0.65	1.1	0.6836	0.5902	0.6272	**0.75**	**1.1**	**0.6879**	**0.7045**	**0.6959**
0.65	1.2	0.6817	0.2130	0.1194	0.75	1.2	0.6811	0.5661	0.6091
0.65	1.3	0.6997	0.2984	0.2897	0.75	1.3	0.6802	0.2697	0.2426
0.70	0.8	0.6793	0.4650	0.5199	0.80	0.8	0.6879	0.2431	0.1865
0.70	0.9	0.6843	0.6353	0.6571	0.80	0.9	0.6961	0.2562	0.2118
0.70	1.0	0.6839	0.6297	0.6535	0.80	1.0	0.6949	0.1962	0.0770
0.70	1.1	0.7038	0.1964	0.0772	0.80	1.1	0.6928	0.1990	0.0842
0.70	1.2	0.6898	0.5461	0.5943	0.80	1.2	0.6910	0.2608	0.2223
0.70	1.3	0.6853	0.5455	0.5934	0.80	1.3	0.6683	0.3127	0.3241

Table C.1
Grid search results of learning rate on CMU-MOSI.

lr	MAE	Corr	Acc_7	Acc_5	Acc_2	F1
0.1	1.5084	0.0910	0.1545	0.1545	0.5938	0.4223
0.05	1.5149	0.0751	0.1545	0.1545	0.4223	0.5938
0.02	1.3138	0.3165	0.2347	0.2347	0.6296	0.6282
0.01	1.0860	0.6380	0.2930	0.3251	0.7504	0.7515
0.005	1.1042	0.5976	0.2930	0.3484	0.7316	0.7332
0.002	1.0421	0.6035	0.3222	0.3673	0.7439	0.7427
0.001	**1.0296**	**0.6591**	**0.3032**	**0.3746**	**0.7591**	**0.7577**
0.0005	1.0699	0.6179	0.3090	0.3090	0.7424	0.7408
0.0002	1.1358	0.5776	0.2726	0.3542	0.7241	0.7227
0.0001	1.3196	0.4177	0.2449	0.2843	0.6296	0.6328

Table C.2
Grid search results of learning rate on CMU-MOSEI.

lr	MAE	Corr	Acc_7	Acc_5	Acc_2	F1
0.1	0.8406	0.0212	0.4137	0.4137	0.6285	0.7718
0.05	0.8377	0.1423	0.4137	0.4137	0.6285	0.7718
0.02	0.7819	0.3327	0.4236	0.4236	0.6715	0.6660
0.01	0.6300	0.6379	0.4911	0.5025	0.7898	0.7955
0.005	0.6433	0.6303	0.4768	0.4885	0.7967	0.7925
0.002	0.6276	0.6463	0.4949	0.5092	0.8064	0.8061
0.001	**0.6331**	**0.6608**	**0.4887**	**0.5061**	**0.8050**	**0.8037**
0.0005	0.6790	0.6186	0.4730	0.4924	0.8003	0.8023
0.0002	0.7198	0.6166	0.4514	0.4736	0.7945	0.7957
0.0001	0.7636	0.4600	0.4131	0.4135	0.7342	0.7268

Table C.3
Grid search results of learning rate on IEMOCAP.

lr	Neutral_F1	Neutral_Acc	Happy_F1	Happy_Acc	Sad_F1	Sad_Acc	Angry_F1	Angry_Acc
0.1	0.2368	0.4083	0.7897	0.8561	0.7032	0.7942	0.6537	0.7580
0.05	0.4399	0.5917	0.7897	0.8561	0.7032	0.7942	0.6537	0.7580
0.02	0.6134	0.6194	0.7897	0.8561	0.7032	0.7942	0.6537	0.7580
0.01	0.6825	0.6844	0.8480	0.8635	0.7848	0.8017	0.8516	0.8507
0.005	0.6848	0.6855	0.8309	0.8667	0.8098	0.8134	0.8719	0.8699
0.002	**0.6831**	**0.6876**	**0.8433**	**0.8625**	**0.8337**	**0.8262**	**0.8647**	**0.8646**
0.001	0.6706	0.6706	0.8355	0.8614	0.7903	0.7825	0.8479	0.8561
0.0005	0.6709	0.6759	0.8345	0.8614	0.8135	0.8113	0.8502	0.8475
0.0002	0.6694	0.6738	0.8195	0.8593	0.7881	0.7910	0.8363	0.8358
0.0001	0.6389	0.6450	0.8016	0.8603	0.7478	0.8017	0.7988	0.8102

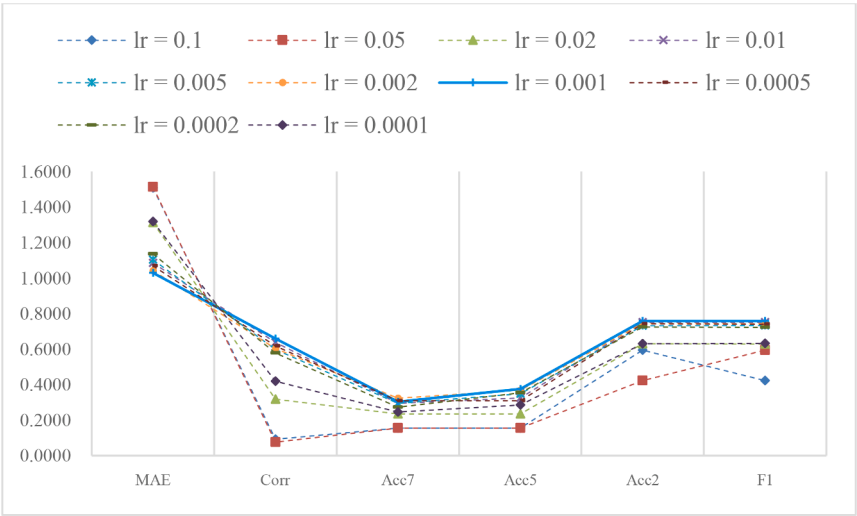


Fig. C.1. Grid search results of learning rate on CMU-MOSI.

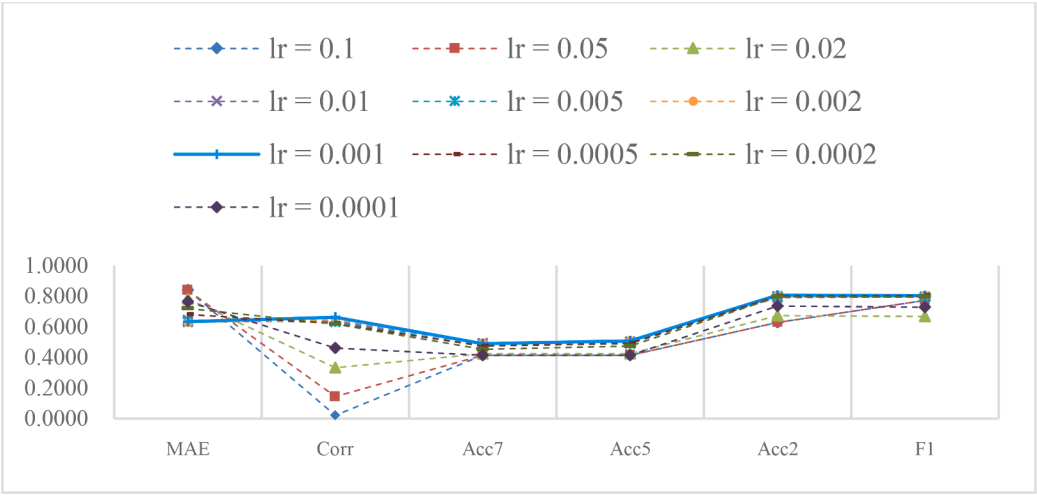


Fig. C.2. Grid search results of learning rate on CMU-MOSEI.

5. Conclusions and future work

In this research, we propose a multimodal time-series method (MTM) for gifting prediction. We extract the features contained in visual, acoustic, textual and numeric information in live streaming setting, and utilize multimodal deep learning methods to predict

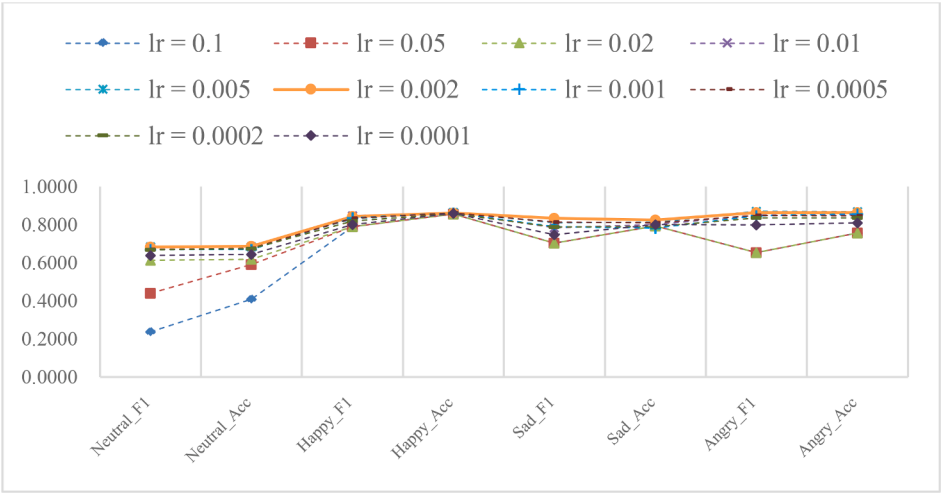


Fig. C.3. Grid search results of learning rate on IEMOCAP.

Table D.1
Grid search results of dimensions on CMU-MOSI.

d	MAE	Corr	Acc_7	Acc_5	Acc_2	F1
120	1.1234	0.5489	0.2828	0.3324	0.7149	0.7135
100	1.1355	0.5776	0.2770	0.3294	0.7256	0.7240
80	1.1381	0.5808	0.2843	0.3367	0.7302	0.7286
60	1.0985	0.5715	0.3047	0.3542	0.7363	0.7349
40	**1.0296**	**0.6591**	**0.3032**	**0.3746**	**0.7591**	**0.7577**
20	1.1815	0.5100	0.2668	0.3105	0.7104	0.7087
10	1.2097	0.4908	0.2697	0.3090	0.6814	0.6796

Table D.2
Grid search results of dimensions on CMU-MOSEI.

d	MAE	Corr	Acc_7	Acc_5	Acc_2	F1
120	0.6576	0.6335	0.4814	0.4947	0.7798	0.7885
100	0.6501	0.6297	0.4768	0.4889	0.7892	0.7895
80	0.6642	0.6082	0.4751	0.4902	0.7914	0.7950
60	0.6443	0.6378	0.4952	0.5128	0.7928	0.7927
40	**0.6331**	**0.6608**	**0.4887**	**0.5061**	**0.8050**	**0.8037**
20	0.6481	0.6292	0.4876	0.5014	0.7931	0.8008
10	0.6587	0.6250	0.4740	0.4846	0.7953	0.7998

Table D.3
Grid search results of dimensions on IEMOCAP.

d	Neutral_F1	Neutral_Acc	Happy_F1	Happy_Acc	Sad_F1	Sad_Acc	Angry_F1	Angry_Acc
120	0.6423	0.6535	0.8098	0.8625	0.7596	0.7825	0.7613	0.7846
100	0.6355	0.6375	0.8277	0.8561	0.7988	0.8060	0.8344	0.8369
80	0.6488	0.6461	0.8391	0.8529	0.8279	0.8284	0.8497	0.8593
60	0.6697	0.6674	0.8369	0.8561	0.8216	0.8230	0.8603	0.8646
40	**0.6831**	**0.6876**	**0.8433**	**0.8625**	**0.8337**	**0.8262**	**0.8647**	**0.8646**
20	0.6882	0.6908	0.8425	0.8646	0.8229	0.8209	0.8666	0.8678
10	0.6612	0.6727	0.8328	0.8678	0.7946	0.7900	0.8534	0.8539

viewers' gifting behavior in real time. As the core part of MTM, multimodal time-series analysis (MTA) module can handle multimodal information effectively. In MTA, an orthogonal projection model is designed to help the target modality gain more information from other modalities. This model can prevent the problem of information redundancy and can easily be combined with cross attention layers. We introduce the joint representation layer to help cross-modal information interaction within the same level, which can be stacked multiple times. We also add residual connections to these joint representation layers to reduce the loss of information at all

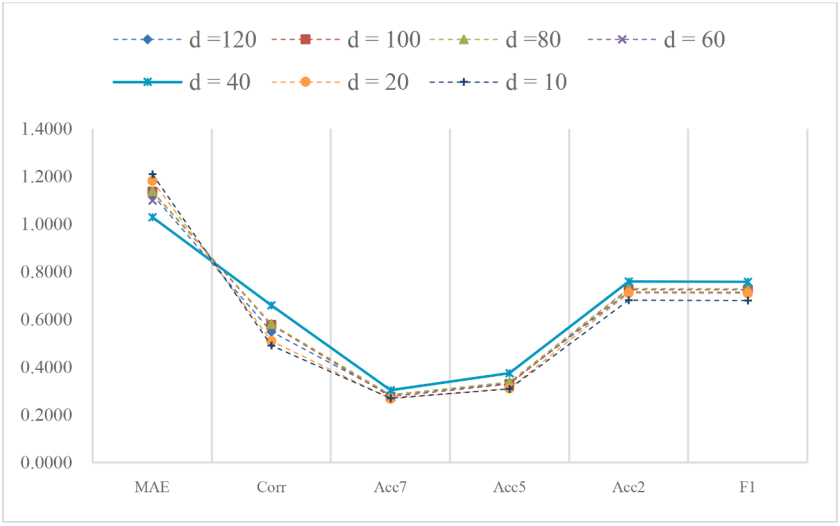


Fig. D.1. Grid search results of dimensions on CMU-MOSI.

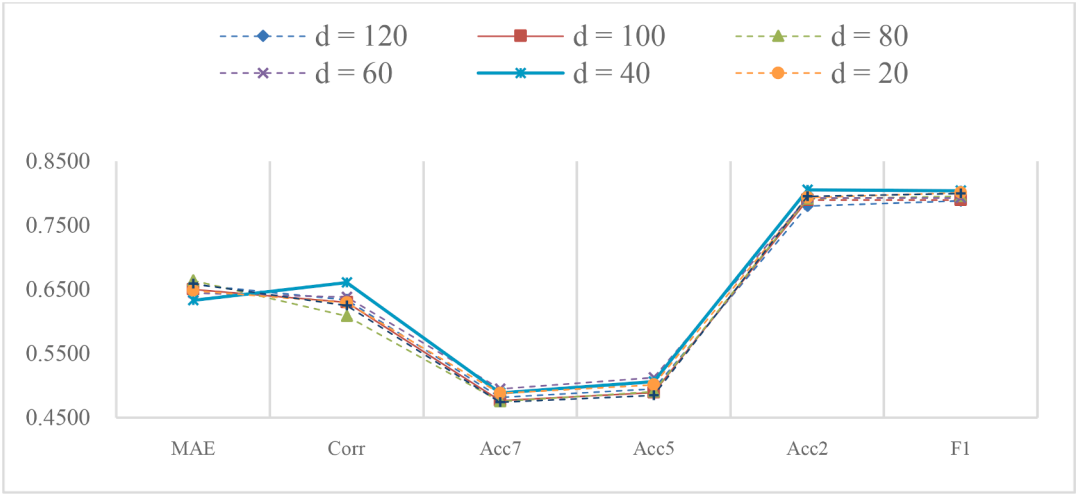


Fig. D.2. Grid search results of dimensions on CMU-MOSEI.

levels. Our model outperforms other advanced baselines by at least 8% on the F1 on our dataset with a 5 min observation window, and leads by 16.54% on the dataset with a 10 min observation window. While achieving improved performance, our model also reasonably controls the growth of the three structure related metrics of parameters, FLOPs, and training time. This makes it suitable for real-time prediction scenarios where timeliness is a requirement. Besides our own dataset, we also verify the robustness and transferability of MTA on three standard datasets of different tasks. Our model cannot only help the streamers to monitor gifting in real time, but also assist the live streaming platforms to obtain more revenue by timely identifying the potential streamers, who may receive more viewer gifts in the following period of time. The dynamic capture of the gifting of live streaming clips also provides a new perspective for the recommendation systems of live streaming platforms.

However, there are inevitably some limitations in our work. Our model relies heavily on modality integrity, which makes the joint representation layer vulnerable to the vacancy of modality. In addition, the way we deal with numeric information has not been fully optimized and needs to be further integrated with the main modalities (visual, acoustic and textual modality). In the future, we hope to design a fusion method that is suitable for all modes and robust for vacant modality. We also plan to provide a larger and better dataset of live streaming for the multimodal deep learning community.

Author statement

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work; there is no professional or other personal interest of any nature or kind in any product, service and/or company that

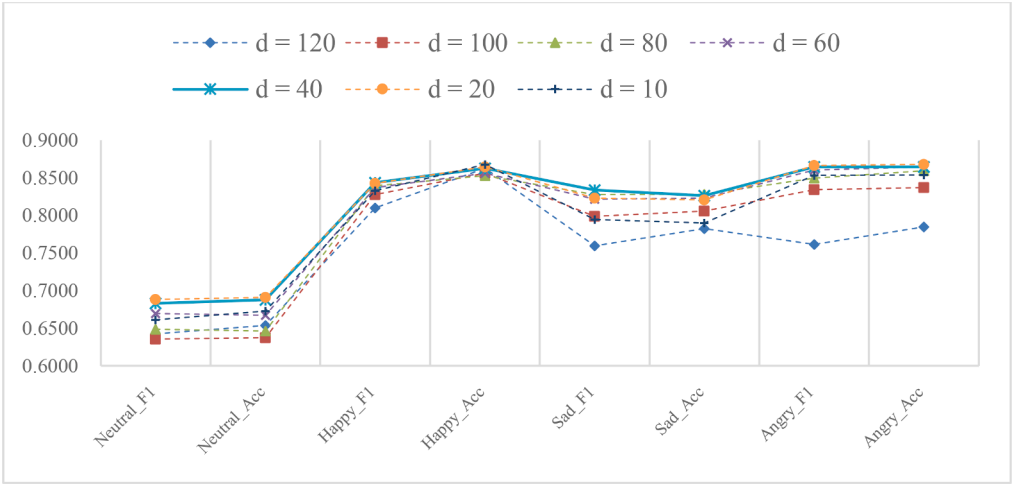
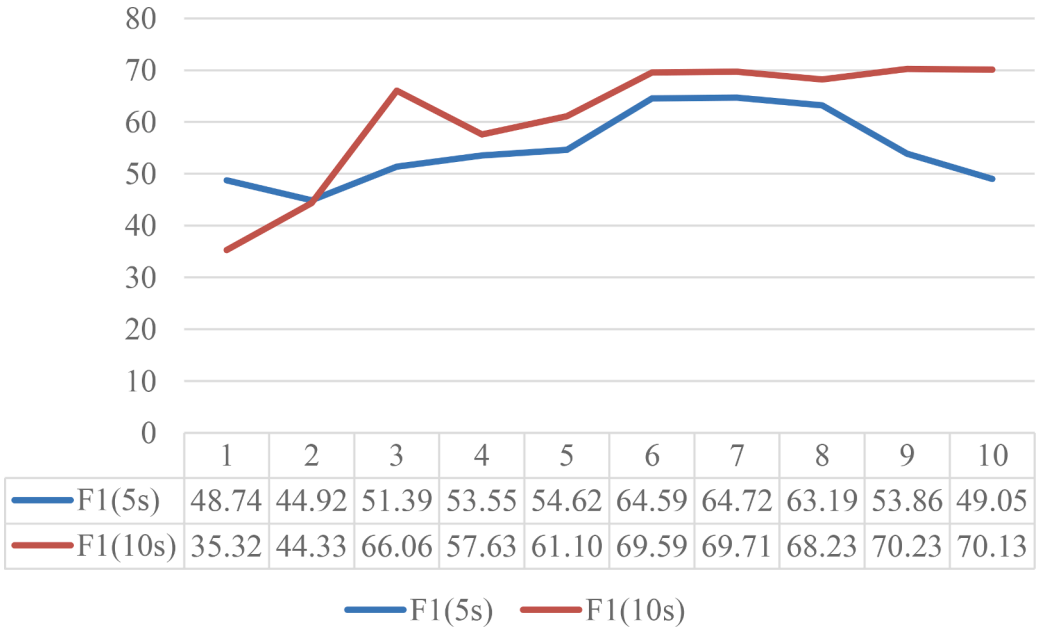


Fig. D.3. Grid search results of dimensions on IEMOCAP.



	1	2	3	4	5	6	7	8	9	10
F1(5s)	48.74	44.92	51.39	53.55	54.62	64.59	64.72	63.19	53.86	49.05
F1(10s)	35.32	44.33	66.06	57.63	61.10	69.59	69.71	68.23	70.23	70.13

Fig. E.1. Grid search results of F1 score under different settings of N.

could be construed as influencing the position presented in, or the review of, the manuscript entitled.

CRediT authorship contribution statement

Dinghao Xi: Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft. **Liumin Tang:** Data curation, Investigation, Formal analysis, Writing – original draft, Methodology, Visualization. **Runyu Chen:** Data curation, Investigation, Formal analysis, Writing – review & editing, Supervision. **Wei Xu:** Conceptualization, Funding acquisition, Investigation, Formal analysis, Methodology, Project administration, Supervision, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 72271233, 72201061, 71771212, and 62172094), Fundamental Research Funds for the Central Universities in UIBE (CXTD12-04), Interdisciplinary-Innovative Research Program of School of Interdisciplinary Studies, Renmin University of China, and School of Interdisciplinary Studies, Renmin University of China.

Appendix A

This appendix gives an example of the OP model's role in high and low redundancy information. We consider the following live streaming scenario: a host dances accompanied by music to attract viewers to watch and tip. Meanwhile, the audience makes live comments about the host's dance, actions, and clothing, but is not interested in the background music. In this case, the information redundancy between the text modality (i.e., viewer comments) and the visual modality (i.e., live streaming video) is high, while the overlap between the text modality and the auditory modality (i.e., background music) is low.

In the subsequent multi-modal information fusion process, our OP model begins to play a role. It will enrich its own information by referencing other modal information while retaining the essence of the target modal information. If the redundancy between other modal information and the target modal information is high, duplicate information will be eliminated. If the redundancy of other modal information with the target modal information is low, more supplementary information from other modalities will be added. An example of OP model's role in high and low redundancy information is shown in Fig. A.1.

Appendix B

This appendix shows how we determine the hyperparameter settings of focal loss. Focal loss consists of two key factors and the main target is to deal with the class imbalance problem. As formula (1) demonstrates, α refers to the weighting factor for class 1 and class 0. According to Lin et al. (2017), α is often set by inverse class frequency. Since the ratio of positive and negative samples of our dataset is about 1:3 (0.25: 0.75), we set search range of α from 0.65 to 0.8. As for the focusing parameter γ, smoothly adjusting the rate at which easy examples are down-weighted. This $(1 - p_t)^{\gamma}$ is often referred to as the modulating factor. The predicted value of the model (p_t) is close to 1, indicating that this sample is easy to train. On the contrary, when p_t tends to 0, it means that the sample is difficult to train. Thus, we can tune the focusing parameter γ to down-weight easy examples (negative samples) and focus training on hard examples. Here we set the grid search range for γ to be 0.8 to 1.3. Values outside this range were also tested but the results were not so satisfactory.

$$\mathrm{FL}(p_t) = -\alpha_t(1 - p_t)^{\gamma}\log(p_t) \tag{1}$$

Here are the grid search results of the focal loss parameters on our dataset (window=5).

Here are the grid search results of the focal loss parameters on our dataset (window=10).

As can be seen from the tables and figures, the darker red parts of the image correspond to the best results we found. This is our process and method for determining the hyperparameters of the focal loss (Figs. B.1 and B.2 Tables B.1 and B.2).

Appendix C

This appendix gives the results of grid search on learning rates. All our hyperparameter settings are determined based on past research and grid experiments. Regarding the setting of learning rates, we take the three standard datasets CMU-MOSI, CMU-MOSEI and IEMOCAP as examples to explain the reasons for our selection. Following the practice of Ghosal et al. (2018) and Tsai et al. (2019), we perform basic grid searches for each hyperparameter and report the average performance of N (N≥5) runs for all our experiments. In practice, we chose N=20 to keep the results stable and reduce the influence of random errors. Here, Table C.1-C.3 report the numeric results of each evaluation indicators under different learning rates (lr), while Figs. C.1–C.3 give a more intuitive visualization of the results.

As can be seen from all the results above, for the CMU-MOSI and CMU-MOSEI datasets, the learning rate setting of 1e-3 can achieve overall better performance in F1-score, accuracy, correlation with relatively low MAE, while learning rate is equal to 2e-3 suits the IEMOCAP dataset. The attributes and distribution of different datasets lead to different network convergence speed and performance. That's how and why we chose different learning rates.

Appendix D

This appendix gives justifications about the chosen of the dimension of embedding layers. Like other hyperparameters, we also perform grid experiments on the dimensions of the model's embeddings layer to get the best setting. Meanwhile, we consider that if the dimension of the hidden layer is too high, it will lead to the overfitting problem of the network. Thus, we set the maximum of the dimension to 120, which is also limited by the GPU memory (RTX 3090, 24GB). Besides, the dimension should be the integer multiple of the numbers of attention head. Thus, we chose d = 120, 100, 80, 60, 40, 20, 10 as our search scope. Here, Tables D.1–D.3 report the numeric results of each evaluation indicators under different dimensions, while Figs. D.1–D.3 give a more intuitive visualization of the

results.

Here are the grid search results of the number of dimensions on CMU-MOSI dataset.

Here are the grid search results of the number of dimensions on CMU-MOSEI dataset.

Here are the grid search results of the number of dimensions on IEMOCAP dataset.

As can be seen from all the results above, the choice of d = 40 can achieve overall better performance with fewer parameters while ensuring the effect. The number of dimensions of the model's embeddings layer is also not the bigger the better, which may cause overfitting problems with big dimensions.

Appendix E

This appendix gives the results of grid search on the number (layers) of crossmodal blocks. Here we give the f1-score under different N settings, respectively on our two datasets (window = 5 and 10 min).

As shown in the Fig. 1, when $N = 6$, the model can achieve satisfactory F1-score. The performance deteriorates when n is decreased, and there is no significant improvement when N is increased. Therefore, choosing N=6 can avoid excessive parameters and achieve good performance Fig. (E.1).

References

Abdu, S. A., Yousef, A. H., & Salem, A. (2021). Multimodal video sentiment analysis using deep learning approaches, a survey. *Information Fusion, 76*, 204–226.

Ariav, I., & Cohen, I. (2019). An end-to-end multimodal voice activity detection using wavenet encoder and residual networks. *IEEE Journal of Selected Topics in Signal Processing, 13*(2), 265–274.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460.

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443.

Ben-Younes, H., Cadene, R., Cord, M., & Thome, N. (2017). Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2612–2620).

Ben-Younes, H., Cadene, R., Thome, N., & Cord, M. (2019). Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *, 33. Proceedings of the AAAI conference on artificial intelligence* (pp. 8102–8109).

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., & Veit, A. (2021). Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10231–10241).

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation, 42*(4), 335–359.

Chen, W. K., Chen, L. S., & Pan, Y. T. (2021). A text mining-based framework to discover the important factors in text reviews for predicting the views of live streaming. *Applied Soft Computing, 111*, Article 107704.

Cui, Y., Yang, Z., & Liu, T. (2022). PERT: Pre-training BERT with permuted language model. arXiv e-prints, arXiv-2203 (pp. 1-14).

D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR), 47*(3), 1–36.

Delbrouck, J.B., Tits, N., Brousmiche, M., & Dupont, S. (2020). A transformer-based joint-encoding for emotion recognition and sentiment analysis. In Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML), Association for Computational Linguistics (pp. 1-7).

Ding, X., Kou, Y., Xu, Y., & Zhang, P. (2022). As uploaders, we have the responsibility": Individualized professionalization of Bilibili uploaders. In *Proceedings of the CHI conference on human factors in computing systems* (pp. 1–14).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the international conference on learning representations* (pp. 1–21).

Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation, 32*(5), 829–864.

Ghosal, D., Akhtar, M. S., Chauhan, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2018). Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3454–3466).

Gönen, M., & Alpaydın, E. (2011). Multiple kernel learning algorithms. *The Journal of Machine Learning Research, 12*, 2211–2268.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, M., Ge, Y., Chen, E., Liu, Q., & Wang, X. (2017). Exploring the emerging type of comment for online videos: Danmu. *ACM Transactions on the Web (TWEB), 12*(1), 1–33.

Hennequin, R., Khlif, A., Voituret, F., & Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software, 5*(50), 2154.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks, 4*(2), 251–257.

James, A. P., & Dasarathy, B. V. (2014). Medical image fusion: A survey of the state of the art. *Information Fusion, 19*, 4–19.

Jan, B., Farman, H., Khan, M., Imran, M., Islam, I. U., Ahmad, A., Ali, S., & Jeon, G. (2019). Deep learning in big data analytics: a comparative study. *Computers & Electrical Engineering, 75*, 275–287.

Jia, A. L., Rao, Y., & Shen, S. (2021). Analyzing and predicting user donations in social live video streaming. In *Proceedings of the IEEE 24th international conference on computer supported cooperative work in design (CSCWD)* (pp. 1256–1261).

Jiang, B., Huang, X., Yang, C., & Yuan, J. (2019). SLTFNet: A spatial and language-temporal tensor fusion network for video moment retrieval. *Information Processing & Management, 56*(6), Article 102104.

Kim, J. H., On, K. W., Lim, W., Kim, J., Ha, J. W., & Zhang, B. T. (2017). Hadamard product for low-rank bilinear pooling. In *Proceedings of the 5th international conference on learning representations* (pp. 1–14).

Lai, H. C., Tsai, J. Y., Shuai, H. H., Huang, J. L., Lee, W. C., & Yang, D. N. (2020). Live multi-streaming and donation recommendations via coupled donation-response tensor factorization. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 665–674).

Lee, G., Jeong, J., Seo, S., Kim, C., & Kang, P. (2018). Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. *Knowledge Based Systems, 152*, 70–82.

Li, R., Lu, Y., Ma, J., & Wang, W. (2021). Examining gifting behavior on live streaming platforms: An identity-based motivation model. *Information & Management, 58*(6), Article 103406.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).
Lin, Y., Yao, D., & Chen, X. (2021). Happiness begets money: Emotion and engagement in live streaming. *Journal of Marketing Research, 58*(3), 417–438.
Liu, H., Tan, K. H., & Pawar, K. (2022). Predicting viewer gifting behavior in sports live streaming platforms: the impact of viewer perception and satisfaction. *Journal of Business Research, 144*, 599–613.
Lu, Z., Xia, H., Heo, S., & Wigdor, D. (2018). You watch, you give, and you engage: a study of live streaming practices in China. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–13).
Murphy, R. R. (2019). Computer vision and machine learning in science fiction. *Science Robotics, 4*(30), eaax7421.
Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning* (pp. 689–696).
Pan, D., Yang, Z., Tan, H., Wu, J., & Lin, H. (2022). Dialogue topic extraction as sentence sequence labeling. In *Proceedings of the CCF international conference on natural language processing and Chinese computing* (pp. 252–262).
Perronnin, F., Sánchez, J., & Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *Proceedings of the European conference on computer vision* (pp. 143–156).
Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE, 91*(9), 1306–1326.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536.
Soleymani, M., Pantic, M., & Pun, T. (2011). Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing, 3*(2), 211–223.
Song, C., Ning, N., Zhang, Y., & Wu, B. (2021). A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management, 58*(1), Article 102437.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.
Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 6558–6569).
Tu, W., Yan, C., Yan, Y., Ding, X., & Sun, L. (2018). Who is earning? Understanding and modeling the virtual gifts behavior of users in live streaming economy. In *Proceedings of the IEEE conference on multimedia information processing and retrieval (MIPR)* (pp. 118–123).
Uppal, S., Bhagat, S., Hazarika, D., Majumder, N., Poria, S., Zimmermann, R., & Zadeh, A. (2022). Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion, 77*, 149–171.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (pp. 5998-6008).
Wang, D., Lee, Y. C., & Fu, W. T. (2019). I love the feeling of being on stage, but I become greedy" Exploring the impact of monetary incentives on live streamers' social interactions and streaming content. *Proceedings of the ACM on Human-Computer Interaction, 3*, 1–24. CSCW.
Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., & Yu, P. (2022). Generalizing to unseen domains: a survey on domain generalization. *IEEE Transactions on Knowledge & Data Engineering*, 1–20.
Wehner, N., Seufert, M., Egger-Lampl, S., Gardlo, B., Casas, P., & Schatz, R. (2020). Scoring high: Analysis and prediction of viewer behavior and engagement in the context of 2018 FIFA WC live streaming. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 807–815).
Wohn, D. Y., Freeman, G., & McLaughlin, C. (2018). Explaining viewers' emotional, instrumental, and financial support provision for live streamers. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–13).
Wu, Q., Sang, Y., & Huang, Y. (2019). Danmaku: A new paradigm of social interaction via online videos. *ACM Transactions on Social Computing, 2*(2), 1–24.
Wu, Y., Niu, G., Chen, Z., & Zhang, D. (2022). Purchasing social attention by tipping: Materialism predicts online tipping in live-streaming platform through self-enhancement motive. *Journal of Consumer Behaviour, 21*(3), 468–480.
Xi, D., Xu, W., Chen, R., Zhou, Y., & Yang, Z. (2021). Sending or not? A multimodal framework for Danmaku comment prediction. *Information Processing & Management, 58*(6), Article 102687.
Xiao, S., Chen, G., Zhang, C., & Li, X. (2022). Complementary or substitutive? A novel deep learning method to leverage text-image interactions for multimodal review helpfulness prediction. *Expert Systems with Applications, 208*, Article 118138.
Xu, X. Y., Luo, X. R., Wu, K., & Zhao, W. (2021). Exploring viewer participation in online video game streaming: A mixed-methods approach. *International Journal of Information Management, 58*, Article 102297.
Yang, K., Xu, H., & Gao, K. (2020). Cm-bert: Cross-modal bert for text-audio sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 521–528).
Yang, L., Na, J.-C., & Yu, J. (2022). Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. *Information Processing & Management, 59*(5), Article 103038.
Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2021). Transfer ability of monolingual wav2vec2.0 for low-resource speech recognition. In *Proceedings of the international joint conference on neural networks (IJCNN)* (pp. 1–6).
Yu, Z., Yu, J., Fan, J., & Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 1821–1830).
Yu, Z., Yu, J., Xiang, C., Fan, J., & Tao, D. (2018). Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems, 29*(12), 5947–5959.
Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 2236–2246) (Volume 1: Long Papers).
Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems, 31*(6), 82–88.
Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2008). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39–58.
Zhai, S., Jaitly, N., Ramapuram, J., Busbridge, D., Likhomanenko, T., Cheng, J. Y., Talbott, W., Huang, C., Goh, H., & Susskind, J. M. (2022). Position prediction as an effective pretraining strategy. In *Proceedings of the international conference on machine learning* (pp. 26010–26027).
Zhang, N., Shen, S. L., Zhou, A., & Jin, Y. F. (2021). Application of LSTM approach for modelling stress–strain behaviour of soil. *Applied Soft Computing, 100*, Article 106959.
Zhou, J., Zhou, J., Ding, Y., & Wang, H. (2019). The magic of Danmaku: A social interaction perspective of gift sending on live streaming platforms. *Electronic Commerce Research and Applications, 34*, Article 100815.
Zhu, Q. S., Zhang, J., Zhang, Z. Q., Wu, M. H., Fang, X., & Dai, L. R. (2022). A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition. In *Proceedings of the ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 3174–3178).

Mr. Xi is a Ph.D. candidate at the School of Information, Renmin University of China, majoring in Computer Application Technology. He obtained his bachelor degree from Renmin University of China. His research interests include social media, business intelligence and multimodal deep learning. His research has been published in Information Processing & Management.

Ms. Tang is a master student at the School of Information, Renmin University of China, majoring in Management Science and Engineering. She obtained her bachelor degree in Computer Science and Technology from Renmin University of China. Her research interests include business analysis, social media, and deep learning.

Dr. Chen is an associate professor at School of Information Technology and Management, University of International Business and Economics. He obtained his doctor degree in Computer Application Technology at Renmin University of China, and B.E. degree from Beijing University of Posts and Telecommunications. His research interests include data mining, business analysis, and financial technology. He has published several papers in international journals and conferences, such as Decision Support Systems, Information Processing and Management, Electronic Commerce Research, and Electronic Commerce Research and Applications.

Dr. Xu is a professor at School of Information, Renmin University of China. He obtained his bachelor and master degree in Mathematics at Xi'an Jiaotong University and doctor degree in Management Science at Chinese Academy of Sciences. His research interests include big data analytics, business analytics and decision support systems. He has published over 150 research papers in international journals and conferences, such as Production and Operations Management, Decision Support Systems, European Journal of Operational Research, IEEE Trans., Information Processing and Management, Information Sciences, International Journal of Production Economics, ICIS and IJCAI.